Exhibit 1

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

IN RE GOOGLE GENERATIVE AI COPYRIGHT LITIGATION

Case No. 23-cv-03440-EKL (SVK)

ORDER RE DISCOVERY DISPUTES

Re: Dkt. Nos. 237, 240, 243

Before the Court are three discovery disputes submitted by the Parties over the last month. The Court will address each in turn, without oral argument. Civ. L.R. 7-1(b).

First, however, the Court strongly admonishes the Parties to respect this Court's Standing Order regarding civil discovery disputes, in particular page limitations, proper use of footnotes and the purpose of the summary chart. The Court's Standing Order is to facilitate the Court's review and evaluation of a dispute; failure to comply is disrespectful of the Court's time and efforts in this endeavor. In particular as it relates to the disputes at hand, the Court does not consider argument in footnotes or argument in the summary chart that is not addressed in the joint statement. Additionally, the page limitation is a maximum, not a minimum, requirement. Both sides' presentations in all three submissions would have benefited from conscientious editing.

I. Dispute re Datasets (Dkt. 237)

Discovery from the datasets used by Google to train the at-issue AI models was the subject of an early compromise proposed by Google, agreed to by Plaintiffs and approved by the Court as a means of managing the vast universe of data underlying Plaintiffs' claims. Order re Discovery Disputes ("June 18 Order") at 2; *see also* Dkt. 161 at 25:23-31:24 (discussion of proposed production and review protocol therefore). Relevant for this discussion, the initial compromise was for Plaintiffs to select 19 exemplar datasets used to train the at-issue models and to which Plaintiffs have since had access for their desired purposes. *See* Dkt. 238-3; Dkt. 161 at

25:23-26:8.[1]

Plaintiffs now seek production of two types of additional datasets. The first consists of five datasets ([redacted], FineWeb, FineWeb2, Fine Web-edu and Common Crawl) that Plaintiffs claim "Google acquired and ingested for training" the at-issue models.[2] Dkt. 238-2 at 3; *see also* Dkt. 238-3 at 2. The second type consists of three additional datasets (The Pile, RedPajama, and RedPajama2) for which Plaintiffs do not have evidence of ingestion by Google. Dkt. 238-2 at 6.

Regarding the first set of the five datasets, the evidence upon which Plaintiffs rest is testimony by Google employee Dr. Xiao that these datasets were ingested by the Core Data Acquisition team ("CDA") and made available for AI training. Dkt. 238-2 at 5; Dkt. 238-4 (Plaintiffs' Excerpts of Xiao Deposition). From this testimony, Plaintiffs argue that their theory of the case "includes the ingestion process as the first step in the infringement" and that they are entitled to discovery on the "whole chain of infringement from the entire training process." As for the second set of three datasets, Plaintiffs acknowledge Dr. Xiao's testimony that the CDA team did not ingest these datasets and rely only on his testimony that he could not confirm that Google did not use these datasets in training. Dkt. 238-2 at 6.

Google argues that, except for [redacted], the seven remaining datasets (across both types identified above) were never used to train the at-issue models. Dkt. 238-2 at 8. Google further argues that Plaintiffs' "ingestion" theory as to the five datasets is not "use in training" as required by the operative complaint and Judge Lee's order "requiring as an 'objective criteria' that a work 'was used by Google to train' its models." *Id*. at 8-9 (quoting Dkt. 128 at 5). As for [redacted], Google acknowledges its use in training Gemini models, stating that its use was identified in Google productions, but Plaintiffs did not select it as an exemplar dataset. *Id*. at 10. Plaintiffs do not dispute this point.

Plaintiffs' request for production of the disputed datasets is **DENIED**. To ascertain the

[1] Discovery flowing from this compromise has not been without issues, some of which have been brought back to the Court for further adjudication.

[2] The Parties, and the deponent Dr. Xiao, use the term "ingest," which, though not defined, in this context appears to mean "acquire." Accordingly, the Court uses the term "ingest" with this understanding.

parameters of relevance and proportionality mandated by Federal Rule of Civil Procedure 26, the Court looks to the Second Amended Complaint ("SAC") which defines the operative class as:

> All persons or entities domiciled in the United States who owned a United States copyright in any work *used by Google to train Google's Generative AI Models* during the Class Period.

Dkt. 234, ¶ 163 (emphasis added). Dr. Xiao, whose deposition is the keystone of Plaintiffs' argument, expressly testified that he did not know if any of the five ingested datasets were used for training. Dkt. 238-4 at 149:3-7; Dkt. 238-5 (Defendant's Excerpts of Xiao Deposition) at 151:1-7. Thus, the Xiao testimony does not support Plaintiffs' objective. Plaintiffs' reliance on Xiao's lack of knowledge as to whether the three remaining datasets were either ingested or used in training fares even worse. There simply is no evidence in the record that the disputed datasets, except for [redacted], were used to train the at-issue models.

Plaintiffs' attempt to bolster the Xiao testimony by arguing that "ingestion" of a dataset alone is sufficient to bring the dataset within the parameters of Rule 26 is not persuasive on the facts of this case. Essentially the argument boils down to "ingestion of data" equals "use in training." However, Plaintiffs' position is undermined by the operative class definition in the SAC, by Judge Lee's requirement for "an objective criteria," and by the agreed-upon bounds within which discovery related to the datasets has been conducted thus far. Datasets used to train the at-issue models were identified as part of the compromise proposed by Google and agreed to by Plaintiffs, and from this pool of datasets Plaintiffs selected 19 exemplar sets for access. Now Plaintiffs want access to additional datasets, possibly acquired by Google but not used to train the models, because, according to Plaintiffs ingesting is the first step in training. While it may be a fair statement that all datasets used to train the models were first ingested, there is no evidence that all ingested models were used to train, as required by the SAC. Accordingly, mere "ingestion" is not sufficient. As for [redacted], though it was used to train some of the at-issue models, Plaintiffs do not, presumably because they cannot, deny that they were aware of [redacted] as a training dataset and did not select it as an exemplar. Accordingly, a request now for production of [redacted] is untimely.

////

II. Dispute re Plaintiffs' Emails (Dkt. 240)

Google seeks sanctions for Plaintiffs' late production of all email addresses used in connection with Google user-accounts, the production of which was previously ordered by this Court. Dkt. 240 at 1. The addresses enable Google to investigate whether Plaintiffs have licensed their works to Google to develop and operate new services. *Id.* In sum, Google complains that Plaintiffs collectively have failed to timely disclose 47 email addresses, all of which, Google contends, should have come to light pursuant to a reasonable investigation which Plaintiffs were required to execute pursuant to this Court's prior Order. *Id.* Google seeks exclusion of certain evidence from Plaintiff McLennan and declarations from counsel and Plaintiffs detailing the search for responsive email addresses. *Id*. at 2.

The Court **DENIES** Google's request. The Court does not find that the request for sanctions as to McLennan is justified given (1) that the relevant blog post, made via a late-disclosed user-account, was made more than 10 years ago, rendering McLennan's alleged memory lapse plausible; coupled with the fact that (2) once the user-account was identified, Google was able to obtain the evidence it claims supports a license. *See* Dkt. 240 at 4. Moreover, Plaintiffs indicated McLennan would be made available for further deposition as to the post and purported license. Accordingly, any prejudice to Google is de minimis. As for the larger ask for declarations from Plaintiffs and from counsel detailing their investigations for all Google user-accounts, the request is overbroad. However, the Court is concerned about Plaintiffs' efforts to comply with this Court's previous Order as evidenced by the apparent ease with which at least some of the addresses could have been identified in a timely manner. Dkt. 240 at 2, 5. The Court is also concerned by Plaintiffs' insupportable suggestion that, despite this Court's prior Order, the discovery was cumulative, duplicative and equally available to Google. *See id.* at 7. To be clear, that argument is soundly rejected. Accordingly, the Court **ORDERS** as follows: No later than **November 17, 2025**, Plaintiffs' counsel is to provide declaration affirming 1) that a thorough and diligent investigation for each Plaintiffs' Google's user-accounts has been conducted; 2) such investigation is now complete; and 3) all user-accounts responsive to this Court's prior Order have been identified to Google.

////

III. Dispute re Google Clawback of Certain Documents (Dkt. 243)

Plaintiffs challenge Google's clawback, either in whole or in part, of 16 documents on the basis of attorney-client privilege.[3] In addition to the Joint Statement, the Parties provide an excerpted privilege log reflecting the disputed documents. Dkt. 244-3. In the first instance, the Court notes that privilege log comports with this Court's requirements in its Standing Order, (*see* Civil and Discovery Referral Matters Standing Order, § 7.c.), including a description of the document setting forth the subject matter addressed and the purpose for which the document was prepared, as well as identifying the attorneys whose legal advice is the basis for the assertion of privilege. *Id*. Plaintiffs challenge clawback on three grounds, each of which is addressed below.

1. Willfulness Defense

Plaintiffs assert that Google is withholding documents that bear on its state of mind, yet "contests allegations that it acted willfully in infringing copyrighted works." Dkt. 244-2 at 2. Specifically, Plaintiffs cite a deposition transcript to demonstrate that Google's counsel instructed a witness not to answer questions regarding Google's copyright compliance where the understanding is based on legal advice. *Id.* at 3. From this foundation, Plaintiffs argue that Google is improperly asserting privilege as a "sword and shield" and ask the Court to order Google to confirm that it will not raise any defenses related to Google's mental state or to produce all documents related to its belief in the lawfulness of its conduct. *Id.* at 4.

The Joint Submission pre-dates Google's operative answer (Dkt. 262 ("Answer")) and (obviously) the deposition relied upon by Plaintiffs. In the Answer, filed on October 16, 2025 Google asserts 17 affirmative defenses, none of which turn on Google's state of mind. *Id.* at 18-21.[4] Accordingly, Plaintiffs' request is unfounded or, at best, premature. Thus, on the current

[3] The Court is informed that the noun is a compound word, "clawback." When used in a verb phrase, it is proper to separate the words as in "clawed back." This is the formatting the Court has adopted.

[4] Google's affirmative defenses are: (1) failure to state a claim generally, (2) license, (3) fair use, (4) deficient copyright registrations, (5) copyright misuse, (6) unclean hands, (7) estoppel, (9) failure to mitigate damages (Google's defenses are mislabeled, skipping #8, but the Court follows the labeling as it appears to limit confusion), (10) statute of limitations, (11) contractual limitations, (12) de minimis use, (13) copyright invalidity, (14) safe harbor under the Digital Millennium Copyright Act, (15) collateral estoppel, (16) lack of copyright notice, (17) lack of standing and (18) a reservation of rights to assert specific defenses against specific plaintiffs/class members.

record, Plaintiffs' challenge to Google's clawback of privileged documents on the basis of a willfulness defense is **DENIED**.

2. Waiver of Privilege: Clawback not Diligent

Plaintiffs assert that Google waived privilege as to clawed back portions of document XXX6921 and document XXX3042.C by failing to act with diligence after Plaintiffs had relied upon the documents in correspondence and depositions. Dkt. 244-2 at 5.[5] Plaintiffs claim that Google clawed back document XXX6921 46 days after Plaintiffs used the document in deposition and clawed back document XXX3042 over 100 days from when Plaintiffs first brought the document to Google's attention. Plaintiffs properly cite to Federal Rule of Evidence 502(b) to determine waiver, pursuant to section 15(d) of the Parties' protective order. Dkt. 119.

Google points this Court's order of June 18, 2025 which called for the expedited production of documents and provides, "Given that production is expedited, any privileged material produced qualifies to be 'clawed back' under Federal Rule of Evidence 502(b)." June 18 Order at 2. Google then posits that in each instance, once the privileged language in these multi-page documents was brought to its attention, it immediately clawed back the relevant portions.

The Court first considers the circumstances of claw back and assertion of waiver as to each document.

XXX6921

The Parties appear to be in accord that this 13-page document was produced pursuant to this Court's June 18 Order and first introduced in deposition on July 25, 2025. Dkt. 244-4 at 5, 9. According to Google, and unrefuted by Plaintiffs, the witness was asked about the document but not about the clawed back passage or the subject matter of the clawed back

[5] Plaintiffs' statements are muddled as to the extent of the waiver as to each document. Plaintiffs initially claim waiver "over all redactions" to document XXX6921 without specifying if there are redactions other than the claw backed portions. 244-2 at 4-5. As for document XXX3042, Plaintiffs initially claim waiver of "the clawed back portion" but in a footnote state that they are challenging all redactions in the document "because they relate to business not legal advice." *Id.* at 5 n. 6. Then in conclusion, as to both documents, Plaintiffs argue that Google waived the "right to claw back" which implicates just the portions of the documents clawed back by Google. *Id.* at 5-6. The Court notes that the facts relied upon by both Parties in this section of the Joint Submission address only two specific instances of portions of documents being clawed back by Google. *Id.* at 5, 9. Accordingly, in this section of this Order, the Court addresses only the propriety of these two instances of clawback and nothing more.

passage. *Id.* at 9. The document was used again in deposition of another witness on September 9, 2025, with questions directed specifically to comments from Google in-house counsel Tom Lue. *Id.* at 5, 9. Google asserts that it immediately clawed back the portion of the document that reflected these comments. *Id.* at 9.

XXX3042

This 11-page document was first referenced by Plaintiffs in a May 11, 2025 letter and later marked as an exhibit at a July[6] deposition.[7] According to Plaintiffs, the May letter cited the document "extensively." Dkt. 244-2 at 5. Google responds that neither the May 11 letter nor the questions at deposition referenced the portion of the document which Google has now clawed back. *Id.* at 9. Neither Party refutes the other's characterizations of the disclosures. Google further claims that it was in preparing its privilege log that it discovered an unredacted fragment of a sentence, which it corrected by clawing back the entire sentence. *Id.* Plaintiffs identify September 2, 2025 as the date Google clawed back "part of this document." *Id.* at 5.

In sum, Plaintiffs argue that diligence should be assessed from the time when the document was brought to Google's attention; Google argues to assess diligence from the time when the specific privileged language, which is the only portion Google is clawing back, was brought to its attention. Interestingly, Rule 502(b) is silent on this distinction, providing in relevant part that "the disclosure does not operate as a waiver" if "the holder promptly took reasonable steps to rectify the error." Courts have construed Rule 502(b) to require action once a party is put on actual or constructive notice of the privileged material—"*i.e.*, when they should know—that they may have produced a privileged communication." *Xu v. FibroGen, Inc.*, 2023 WL 3475722, at *3 (N.D. Cal. May 15, 2023). For constructive notice, courts "look to when the party, acting with reasonable diligence, should have discovered the inadvertent disclosure." *Id.* (citation omitted).

////

[6] Plaintiffs cite to a "30(b)6" deposition on July 2, 2025; Google cites to a "corporate" deposition on July 25, 2025. Dkt. 244-2 at 5, 9. This discrepancy is not material to the Court's analysis.

[7] Given the reference to the document in a May letter, the Court concludes this document was not produced pursuant to its June 18 Order. Google is silent as to the date of production.

XXX6921

Whether clawback of this document is timely is a close call. However, in light of the circumstances of production of this document pursuant to this Court's June 18 Order, along with the fact that the privileged portion of the document was not identified or relied upon by Plaintiffs until the deposition of September 9, 2025, the Court finds that Google's clawback on September 9, 2025 was timely and therefore Google did not waive its right to assert the privilege.

XXX3042

The circumstances surrounding this document calls for a different result. The production of this document preceded this Court's June 18 Order. The document was twice brought to Google's attention by Plaintiff, including an extensive discussion of the document in May 2025. Further, the document, at 11 pages, is not so voluminous that it would be unreasonable find constructive notice of the entire contents of the document through this notice. Accordingly, the Court finds that Google waived its privilege as to portion of this document clawed back on September 2, 2025.

3. Failure to Establish Privilege

Plaintiffs' final argument appears to be a "catch-all" effort to seek production of all sixteen disputed documents. Dkt. 244-2 at 6. Plaintiffs argue that Google has failed to demonstrate that the primary purpose of both the redacted and withheld documents is for legal advice. *Id.*

Plaintiffs support this argument first by citing to purported "technical records prepared by Google's technical staff" on the privilege log, (Dkt. 244-3), concluding that these documents' "primary function is not the procurement of legal advice." *Id*. First, Plaintiffs mischaracterize these log entries. Plaintiffs cite to document XXX3053.C, describing it as a "data card for Gemini v1." *Id.* But the privilege log actually describes the document an "*annotated draft memorandum* on [redacted]" Dkt. 244-3 at 4 (emphasis added). The evidence before the Court is that the disputed document is not a data card but a memorandum discussing a data card. The log also indicates that the document is merely redacted, not withheld, which, as Google points out, would be appropriate for a document that contains a factual technical

discussion separate from legal advice. The additional examples of "technical records" relied upon by Plaintiffs are also logged as "annotated draft memorandum" or "memorandum," refer to legal advice and, significantly, are merely redacted and not withheld. *Id.* at 4-5. Google's handling of these documents containing legal and technical advice is proper.

Plaintiffs also complain that Google has withheld from production documents prepared by non-attorneys, which Plaintiffs urge should be narrowly redacted to protect only legal advice. Dkt. 244-2 at 5. Google correctly points out that documents—even those prepared by non-attorneys—may be privileged in their entirety when they reflect legal advice or were created for or at the direction of counsel. *Id.* at 11 (citing *In re Google RTB Consumer Priv. Litig.*, 2023 WL 1787160, at *4-*5 (N.D. Cal. Feb. 6, 2023). Google further explains that these withheld documents are of two types, notes from "legal sync" meetings addressing legal questions related to technical work and documents prepared at counsel's direction. *Id.* As such, Google posits, the documents are properly withheld in their entirety. Google's privilege log supports Google's position. In contrast, Plaintiffs offer only conclusions and argument unsupported by any facts.

In sum, Plaintiffs challenge on this front—that Google has failed to establish that the primary purpose of the redacted and withheld documents is legal advice—is not only an incorrect statement of the law as to documents which are solely for the purpose of transmitting legal advice but is not supported by the record before the Court and is therefore **DENIED.**

SO ORDERED.

Dated: November 10, 2025

SUSAN VAN KEULEN
United States Magistrate Judge